CSC343 Project: Cleaning Decisions - Chloe Nguyen, Patrick Lin

Note that the entire schema was redesigned in light of the issues pointed out during Phase 2 and phase 3 marking. Hence, this entire section is new.

Changelog:

- Removed tables that were not necessary to the schema.
- Ensured all attributes are relevant to the table's purpose.
- Ensured all tables follow good design patterns.
    - In particular, our tables do not violate BCNF.
- Added a new USHoliday table.

To begin, we took a subset of our original data (NYC crashes data, US holidays data, NYC weather data) to include only information on the year 2017. To split our original datasets into tables for our schema design, we utilized the Python library Pandas to work with CSV files.

Table 1: NYCAccidents(collisionID, crashTime, longitude, latitude)

- We chose to include data with missing information.

Table 2: NYCDailyWeather(day, station, minTemp, maxTemp, avgTemp, precipitation, snowfall, snowDepth, waterSnowOnGround, waterSnowfall, avgWindSpeed)

- We included data with missing attributes. This is because some weather stations report only certain attributes (e.g. a station may report only on temperature or only on precipitation) and we want to retain as much information reported as possible.
- We did not set default values for any attributes with null values. This is because having a null value for a certain attribute does not allow us to infer any information. Consider if we had a null value for the attribute snowfall, for example. We can't simply set the default value to 0 since the report might've come from a weather station that only reported on temperature and not snowfall. Hence the default value of 0 is incorrect since there could've been snowfall that wasn't reported by that certain station.

Table 3: NYCCasualties(collisionID, motoristKilled, motoristInjured, cyclistKilled, cyclistInjured, pedestrianKilled, pedestrianInjured)

- We chose to only include accidents with at least 1 confirmed injury or at death
- We include accidents with missing information on injuries and deaths (of pedestrians, motorists and cyclists) as long as there is at least 1 confirmed injury or death

Table 4: NYCFactors(collisionID, factor1, factor2, factor3, factor4, factor5, vehicle1, vehicle2, vehicle3, vehicle4, vehicle5)

- We chose to include data with missing information.

Table 5: NYCWeatherStations(stationID, longitude, latitude)

- We chose to include data with missing information.

Table 6: USHolidays(day, holiday)

- We did not include data with missing information because tuples cannot have null value for day or holiday attributes since (day, holiday) is a key